International Journal of Wavelets, Multiresolution
and Information Processing
Vol. 21, No. 2 (2023) 2250056 (20 pages)

DOI: 10.1142/S0219691322500564



# Jump detection in high-frequency financial data using wavelets

Mateus Gonzalez de Freitas Pinto
*Institute of Mathematics and Statistics*
*University of São Paulo, São Paulo, Brazil*
*mateusgf@ime.usp.br*

Guilherme de Oliveira Lima C. Marques
*Center for Engineering*
*Modeling and Applied Social Sciences*
*Federal University of ABC, Brazil*
*guilherme.lima@ufabc.edu.br*

Chang Chiann
*Institute of Mathematics and Statistics*
*University of São Paulo, São Paulo, Brazil*
*chang@ime.usp.br*

Received 3 February 2022
Revised 6 November 2022
Accepted 8 November 2022
Published 3 December 2022

The presence of spikes or cusps in high-frequency return series might generate problems in terms of inference and estimation of the parameters in volatility models. For example, the presence of jumps in a time series can influence sample autocorrelations, which can cause misidentification or generate spurious ARCH effects. On the other hand, these jumps might also hide relevant heteroskedastic behavior of the dependence structure of a series, leading to identification issues and a poorer fit to a model. This paper proposes a wavelet-shrinkage method to separate out jumps in high-frequency financial series, fitting a suitable model that accounts for its stylized facts. We also perform simulation studies to assess the effectiveness of the proposed method, in addition to illustrating the effect of the jumps in time series. Lastly, we use the methodology to model real high-frequency time series of stocks traded on the Brazilian Stock Exchange and OTC and a series of cryptocurrencies trades.

*Keywords*: Volatility; wavelets; high-frequency data; jumps; long memory; INGARCH.

## 1. Introduction

High-frequency time series usually contain stylized facts that need to be taken into consideration when modeling their behavior. According to Ref. 13, stylized facts can be defined as a set of statistical properties emerging from independent empirical studies of assets. These peculiarities are partial due to price formation and market microstructure effects, which are usually shadowed in lower frequencies and other types of time series, but also depend on the liquidity of the financial markets in which the assets are traded.[44]

Typical stylized facts, some of which are listed in Refs. 13 and 14, are long-range dependence on the conditional mean and conditional variance, intraday jumps, volatility clusters, fat-tailedness, non-normality, and skewness. In terms of structure, the data is available in the “tick-by-tick” format, which means that the quantities of interest are not usually available in homogeneously spaced time, but rather are irregularly spaced.

The reasons why these stylized facts are more apparent at lower frequencies in financial time series are still unclear, but there is a consensus in the literature that this is related to the effects caused by price formation and market microstructure. For instance, Ref. 14 states that, at higher frequencies, the middle price is subject to microstructure effects, which overshadow the lower-frequency properties of financial series.

Jumps can be seen as outliers in either magnitude or as an intervention in the series, and they have their own dynamics and memory structure. Although these jumps or cusps are generally abnormally big, the magnitude of what constitutes a jump depends on the volatility of the data generating process, mainly when dealing with heteroskedastic processes.

It is well known that outliers can influence sample autocorrelations $\hat{\rho}(k)$ of a time series and this might lead to identification problems, as discussed thoroughly in Refs. 10, 26 and 40. References 11 and 27 mention that the presence of an outlier might lead to several problems in terms of identification, such as bias in the estimation of the autocorrelation function (ACF) and partial autocorrelation function (PACF), and even in the ARMA parameters of a given series.

Situations of which outliers behave as interventions in the time series are discussed in Ref. 9, where they show that it is possible to decompose a contaminated series $Z_t$ into two parts

$$Z_t = Y_t + \eta_t(S,\omega), \tag{1}$$

where $Y_t$ denotes the outlier-free series and $\eta_t(S,\omega)$ an exogenous intervention effect located at $t = S$ with magnitude $\omega$. For instance, one can consider an additive outlier of the following form:

$$\eta_t(S,\omega) = \omega I_t^{(S)}, \tag{2}$$

where $I_t^{(S)}$ denotes the indicator function

$$I_t^{(S)} = \begin{cases} 0 & \text{if } t \neq S, \\ 1 & \text{if } t = S, \end{cases} \tag{3}$$

which means that, in (2), the parameter $S$ indicates where the outlier occurs in time, and $\omega$ indicates the magnitude of this outlier. The author proves that, in this case, the sample ACF will have a limiting behavior of zero whenever the magnitude of an additive outlier increases compared to the other points in the observed time series. Reference 9 states that significant additive outliers could thoroughly wash out the information in the sample ACF, as the behavior of the latter would dominate.

Other problems caused by abnormally large values in the series will show up in the maximum likelihood estimation. For instance, Mendes[33] showed that jumps in the GARCH models would generate biased maximum likelihood estimates for the GARCH parameters. Hotta and Zevallos[28] similarly showed that outliers can either lead to spurious conditional heteroskedasticity or hide ARCH effects, in a manner similar to the ARMA models results presented in Ref. 9. Furthermore, cusps can hide legitimate heteroskedasticity when isolated or significant, which was also identified and discussed in Ref. 33 and have effects in forecasts as in Refs. 8 and 12.

The intraday log-returns of the stocks chosen for our empirical analysis usually display price jumps, some of them due to overnight price changes. As a result, not dealing with these cusps in the series might lead to misidentification and a poorer model fit.

On the other hand, these jumps might not necessarily happen in equally spaced timestamps and have different magnitude from one to the other and a serial dependence structure, which suggests that one could implement a stochastic approach rather than treating them simply as additive outliers. This not only addresses the problem of understanding the distance between jumps and explaining their dynamics, but also deals with their magnitudes over time. Moreover, removing these jumps and not modeling them neglects the fact that they often carry information and are part of the observed series.

In order to perform the modeling, we proposed using wavelet shrinkage in order to identify the jump series and wavelet decomposition to separate out the jumps from the time series before modeling it. As a means of comparison, we also fit a model that does not separate out these jumps from the time series, in order to demonstrate the resulting misidentification and how the model resulted in a poorer fit. We fit two models separately: one for the series after separating out the jumps and another for the jump series. After uncoupling the jumps, we apply long memory models for the conditional mean and conditional variance, which account for its stylized facts. This is consistent with the literature on financial time series. Lastly, we fit an INGARCH model to the jump series after a suitable transformation.

We do so because the jump series has a significant probability mass at zero, is asymmetrical, and the jumps occur with irregular time spacing, with varying magnitude, and are non-negative. Therefore, a Poisson-like process can be used to model these jumps after an appropriate transformation that allows the use of the autoregressive conditional Poisson model to this financial time series.

This paper is structured as follows. In Sec. 2, we provide some backgrounds on wavelets and long-memory models which will be used to model the high-frequency data as well as on the INGARCH model. In Sec. 3, we discuss the data and the methodology used. Section 4 describes simulations performed to illustrate the effects of jumps in time series and assess the effectiveness of the proposed methodology. In Sec. 5, we describe the procedure for modeling asset returns for three different assets negotiated on the Brazilian Stock Exchange, and OTC (B3 S.A. — Brasil, Bolsa, Balcão), sampled every minute, and cryptocurrency traded on the Gemini Exchange. We conclude with some final remarks.



## 2. Backgrounds

### 2.1. *Wavelets*

Wavelets are functions that form a basis to represent all functions of class $\mathcal{L}_2(\mathbb{R})$. One of the advantages of using wavelets over other bases is that wavelets are localized in time and scale. Thus, a discontinuity in a function $f(t)$ will only influence the wavelets that are near it and just those wavelet coefficients associated with the influenced wavelets that overlaps the discontinuity, unlike the Fourier basis.[35] See Refs. 15 and 34 for additional background information on wavelets. For the relation between wavelets and fractal geometry, see Refs. 3, 22–24, 42, 43 and 25.

The wavelet bases are generated from binary re-scales and translates the $\psi$ mother wavelet. Let $\psi_{j,k}(t)$, $j,k \in \mathbb{Z}$ be a family of functions defined as translations and re-scales of a single mother function $\psi(t) \in \mathcal{L}_2(\mathbb{R})$, which is given by[41]

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad j,k \in \mathbb{Z}. \tag{4}$$

In Ref. 32, it is proved that from a mother wavelet $\psi$ and a father wavelet $\phi$, one can form an orthonormal system, setting

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \tag{5}$$

and using Eq. (4), a multiresolution decomposition in $L_2(\mathbb{R})$ can be written as

$$\mathcal{L}_2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j = V_0 \oplus \bigoplus_{j\geq 0} W_j, \tag{6}$$

where $W_j$ is the subspace generated by $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ and $V_0$ is the subspace generated by $\{\phi_{0,k}\}_{k\in\mathbb{Z}}$. Hence, for $f \in \mathcal{L}_2(\mathbb{R})$, we may write uniquely a wavelet decomposition

$$f(t) = c_{0,0}\phi(t) + \sum_{j\geq 0}\sum_{k} d_{j,k}\psi_{j,k}(t), \tag{7}$$

with coefficients given by the inner product

$$d_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt, \tag{8}$$

$$c_{j,k} = \int_{-\infty}^{\infty} f(t)\phi_{j,k}(t)dt. \tag{9}$$

Let $\{X_t, t = 1, 2, \dots T\}$ be a time series with $T = 2^J$ for a given $J \in \mathbb{Z}^+$. The discrete wavelet transform (DWT) is given by

$$X_t = c_{0,0}\phi(t) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t), \tag{10}$$

with coefficients given by the inner product. Consider

$$S_0(t) = c_{0,0}\phi_{0,0}(t), \tag{11}$$

$$D_j(t) = \sum_k d_{j,k}\psi_{j,k}(t). \tag{12}$$

Equation (11) is called smooth signal, whereas Eq. (12) is called detail signal of level $j$. Then, according to Eq. (6), it follows that

$$f(t) \approx S_0(t) + D_0(t) + D_1(t) + \cdots + D_{J-1}(t), \tag{13}$$

the approximation above is called multiresolution decomposition of the signal. From the multiresolution decomposition, one can perform the wavelet shrinkage, which consists on passing a filter through one or several specific levels of the decomposition and then, using an orthonormal family, reconstructing the signal with the corresponding antitransform.[6]

For the shrinkage procedure, in this paper, we use the hard threshold

$$\delta_\lambda^H(t) = \begin{cases} 0 & \text{if } |t| \le \lambda \\ t & \text{if } |t| > \lambda, \end{cases} \tag{14}$$

with the universal policy

$$\lambda = \lambda_{j,T} = \sigma\sqrt{2\log T}. \tag{15}$$

There are several alternative shrinkage functions that are used to generate the $\sigma$ quantity in the universal policy in Eq. (15). The choice is more frequently either the usual $L_p$ norm, the median of absolute deviation from the median (denoted as MAD), or even the usual standard deviation, as in Ref. 6, although the MAD tends to perform better in the presence of outliers. Therefore, we use the MAD function or the $L_2$ norm in our shrinkage procedure. Further discussion of wavelet shrinkage can be found in Refs. 15 and 32.

### 2.2. *Long-memory models*

Long-memory models are used in order to account for long-range dependence, as they are a more parsimonious alternative compared to a short-memory model with several parameters, which are based on the mathematical concept of fractional integration $I(d)$, $d \in \mathbb{R}$. In order to model the long-range dependence of the conditional mean, the ARIMA model has been extended by the autorregressive fractionally integrated moving average model (ARFIMA). Likewise the GARCH model has been extended by the fractionally integrated generalized conditional heteroskedasticity (FIGARCH) models.[2,7,31]

Although, in liquid markets, the log-return series display little or even no serial correlation, squared and absolute log-returns are almost always positively autocorrelated, mainly because of the market microstructure properties.[13]

In Ref. 37, the authors claim that this dependency in the high-frequency financial data is due to price discreteness and that there is often a spread between the price paid by the buyer and seller-initiated trades.

In this paper, we use the ARFIMA model

$$\Phi(B)\Delta^d y_t = \Theta(B)\epsilon_t, \tag{16}$$

where $|d| < 0.5$, $\epsilon_t$ is white noise with zero mean and $\sigma^2$ variance, $B$ is the backshift operator, $\Phi(B)$ and $\Theta(B)$ are, respectively, the autoregressive and moving average polynomials, $\Delta^d = \sum_{k=0}^{\infty} \binom{d}{k}(-1)^k B^k$ and the roots of $\Phi(z) = 0$ and $\Theta(z) = 0$ lie outside the unit circle.[2,5,38]

We also use the FIGARCH model for the long-range dependence in the conditional variance[1]

$$\left(1 - \sum_{j=1}^{n} \beta_j B^j\right) \sigma_t^2 = \alpha + \left(1 - \sum_{j=1}^{n} \beta_j B^j - \left(1 - \sum_{i=1}^{m} \alpha_i B^i\right)(1-B)^d\right) r_t^2, \tag{17}$$

where $r_t = \epsilon_t \sigma_t$, $\epsilon_t \sim D(0, 1)$ and $0 < d < 1$.

As shown in Ref. 1, even if the FIGARCH process in Eq. (17) is not second-order stationary, it is ergodic and strictly stationary. For a general review of FIGARCH models and their properties, see Refs. 19 and 7.

### 2.3. *INGARCH models*

Consider a count time series $\{Y_t, t \in \mathbb{Z}\}$. The goal is to model the conditional mean $E(Y_t \mid \Omega_t) = \lambda_t$ of the time series, where $\Omega_t$ denotes the set of all previous information available, and the linear predictor is given as

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \hat{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}), \tag{18}$$

where $g : \mathbb{R}_+ \to \mathbb{R}$ is the link function as in generalized linear model (GLMs), $\hat{g} : \mathbb{N}_0 \to \mathbb{R}$ is a transformation function. In order to perform the regression on arbitrary past observations, we consider the sets $P = \{i_1, i_2, \dots i_p\}$ and $Q = \{j_1, j_2, \dots j_q\}$, which enables the regression to be made in $Y_{t-i_k}$ and $\lambda_{t-j_l}$ for the given $i_k \in P$ and $j_l \in Q$.[30]

Given the identity link $g(x) = x = \hat{g}(x)$, $P = \{1, 2, \dots p\}$, $Q = \{1, 2, \dots, q\}$ and assuming that $Y_t|\Omega_t \sim \text{Poisson}(\lambda_t)$, then we have the INGARCH$(p, q)$ model, sometimes referred to as autoregressive conditional Poisson models (ACP),[17] with the following form:

$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \beta_k Y_{t-i_k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-j_l}, \tag{19}$$

The model described in Eq. (18) is fitted and estimated using the quasi maximum-likelihood method. For the INGARCH$(p, q)$ model, in order to ensure positivity of the conditional mean, all the parameters must be non-negative and, in order to ensure stationarity and ergodicity of the solution, these conditions must hold: $\sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1$ and $\beta_0 > 0, \beta_1 \cdots \beta_p, \alpha_1 \cdots \alpha_q \geq 0$.[30] For a broader review, see Refs. 18 and 17.

## 3. Methodology

In this paper, we propose the following procedures to model high-frequency financial time series while accounting for the jump effects:

(1) Calculate the log-return series $r_t = 100 \times (\ln(P_t) - \ln(P_{t-1}))$ for $t = 2, 3, \dots T$ from the homogeneous time series and obtain the absolute returns $|r_t|$.
(2) Apply the discrete-time wavelet transform (DTWT) to $|r_t|$ and use the wavelet shrinkage to obtain a "smooth" series, which will indicate the intraday jumps.
(3) Separate the $|r_t|$ series into two parts: the jump series resulting from the wavelet shrinkage procedure ($s_t$) and the return series discounting the jumps identified with wavelets, $x_t = |r_t| - s_t$.
(4) Series $x_t$ is modeled according to an ARFIMA-FI(E)GARCH model to account for the stylized facts usually observed in these high-frequency financial time series
(5) We consider $S_t = \lceil 10 \times s_t \rceil$ and model $S_t$ according to an INGARCH process with Poisson innovations.

In step 2 of the procedure, the Haar wavelet was employed since it performed better empirically in terms of identifying discontinuities as breaks or jumps. A brief discussion of this issue is, for instance, found in Ref. 16.

The Haar mother and father wavelets are, respectively,

$$\psi(t) = \begin{cases} +1 & \text{if } 0 \leq t < \dfrac{1}{2} \\ -1 & \text{if } \dfrac{1}{2} \leq t < 1 \\ 0 & \text{else} \end{cases} \tag{20}$$

and

$$\phi(t) = \mathbb{I}_{[0,1]}(t). \tag{21}$$

The choice of 10 for scaling $S_t$ in step 5 is made to preserve four decimal points precision. The fitted values can be rescaled back to match the original jump series.

We compare the results obtained using the methodology described above with an approach that does not separate out the jumps in the returns. To do so, we use the root mean squared error (RMSE) as means of comparison. For the ARFIMA-GARCH model without separating the jumps, the RMSE is obtained by

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}((r_t - \hat{r}_t)^2 - \hat{\sigma}_t^2)^2}, \tag{22}$$

where $\hat{r}_t$ denotes fitted values of the ARFIMA model, $r_t$ denotes the log-returns and $\hat{\sigma}_t^2$ denotes the estimated conditional variance. For the proposed methodology using wavelet decomposition, the RMSE is calculated using

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}((r_t - \hat{s}_t - \hat{x}_t)^2 - \hat{\sigma}_t^2)^2}, \tag{23}$$

where $x_t = r_t - s_t$ denotes the series after separating the identified jumps using the wavelet procedure, $\hat{x}_t$ denotes fitted value of the ARFIMA model fitted to the $x_t$ series, $\hat{s}_t$ the fitted INGARCH model for the jump series and $\hat{\sigma}_t^2$ the estimated conditional variance. The choice of the RMSE instead of other metrics is due to the fact that the mean is more sensitive to outliers rather than the median.

In times of higher volatility, an abnormal return tends to be bigger than an abnormal return in times of low volatility. This is the main reason that justifies the use of non-parametric estimates, identifying jumps in a less arbitrary form that relies solely on the data. The purpose of the RMSE comparison is to determine if jump removal leads to a more accurate estimate of the volatility.

## 4. Simulation

In order to illustrate the proposed methodology, simulation studies were performed employing FIGARCH and INGARCH models. We consider jumps approximately as a realization of an INGARCH$(1, 1)$ process. We use wavelets to separate the FIGARCH from the INGARCH parts, then model them separately and assess the model's performance when compared to an approach where we do not separate out the jumps from the series, but rather model the series directly using ARFIMA-GARCH models. We intend to show that using this wavelet-based approach improves the accuracy of the estimates significantly in terms of RMSE.

In this simulation, jumps were added to the series in the form of an INGARCH$(1, 1)$ process with Poisson innovations, as in (19) with $p = q = 1$.

It is important to stress that the aim of these simulations is not to verify or discuss any properties of the estimators, but rather to illustrate the effect of jumps of the approximate form of an INGARCH added to an FIGARCH model as well as to show that, when using the procedure described in the methodology section, the fitted model is more accurate in terms of RMSE than a model fitted without removing these jumps.

The simulation procedure will be as follows: first, the FIGARCH series is generated. Then the jumps are added, following

$$\begin{cases} Z_t = Y_t + X_t \\ X_t|\Omega_t \sim \text{Poisson}(\lambda_t) \\ \lambda_t = \beta_0 + \beta_1 X_{t-1} + \alpha_1 \lambda_{t-1} \end{cases}, \tag{24}$$

where $Y_t$ denotes the jump-free FIGARCH series and $X_t$ denotes a realization of the INGARCH$(1,1)$ used to generate the jumps, generated as a Poisson with parameter $\lambda_t$.

We fit a model for the $100 \times Z_t$ series without the jumps separated out and, after that, we consider the estimated jump series $s_t$ using wavelet decomposition and separate out the jumps in the same fashion described in the methodology section. We obtain an estimate for the jump-free series, say $100 \times \hat{X}_t$, and for $100 \times \hat{Y}_t$ and fit models to both series.

The simulated data is multiplied by 100 in order to rescale the data, which improves the performance of the numerical optimization procedure of the maximum-likelihood estimation of the FIGARCH models, but it changes nothing in the coefficients we are interested in, just rescales the data.

We simulated $n = 1000$ replicas with $T = 16384$, the same sample size used in an application with real financial time series, and performed the procedure for each of them. In order to do so, we rely on the choice of the best GARCH model based on the heuristic proposed in Ref. 39: we fit ARCH$(1)$, ARCH$(2)$, GARCH$(1,1)$, GARCH$(1,2)$, GARCH$(2,1)$ and GARCH$(2,2)$ models and select the best one in terms of BIC. The same method was used to choose the conditional mean model: the ARFIMA$(p,d,q)$ order is determined by searching for $p \in \{0,1,2\}$ and $q \in \{0,1,2\}$. Moreover, $d \in (-0.5, 0.5)$, including the possibility of choosing a model where $d = 0$, choosing the value corresponding to the best BIC. Finally, we also compare the two previous models with an ARFIMA$(p,d,q)$-FIGARCH$(1,1)$ model.

We shall discuss the details for only one of the 1000 replicas. Let $Y_t \sim$ FIGARCH$(1,d,1)$, with $\omega = 0.0002$, ARCH$(1) = 0.43$, GARCH$(1) = 0.53$ and $d = 0.17$. Assume also $X_t|\Omega_t \sim \text{Poisson}(\lambda_t)$, with $\lambda_t = \beta_0 + \beta_1 X_{t-1} + \alpha_1 \lambda_{t-1}$, with $\beta_0 = 0.001$, $\beta_1 = 0.51$ and $\alpha_1 = 0.38$. We generate the contaminated series

$$Z_t = Y_t + X_t, \tag{25}$$

as in Fig. 1.

While in Fig. 2(a) the samples ACF and PACF of $Y_t^2$ decay slowly, suggesting the possibility of fitting an FIGARCH model to the conditional variance of the

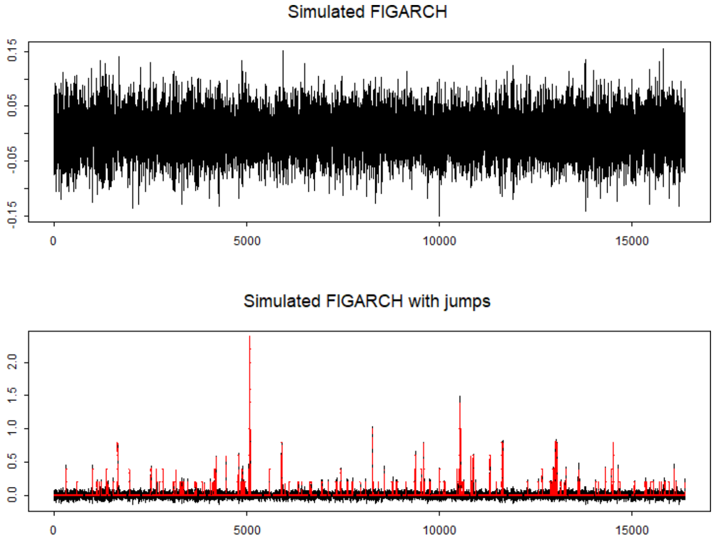
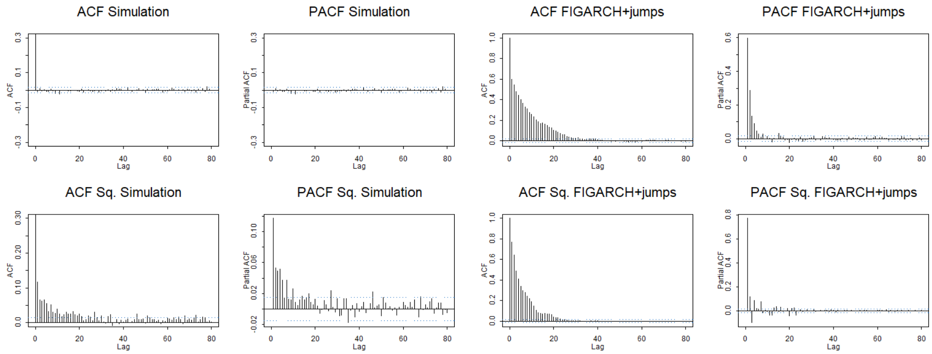



Fig. 1. Simulated FIGARCH series and INGARCH jumps.




(a) (b)

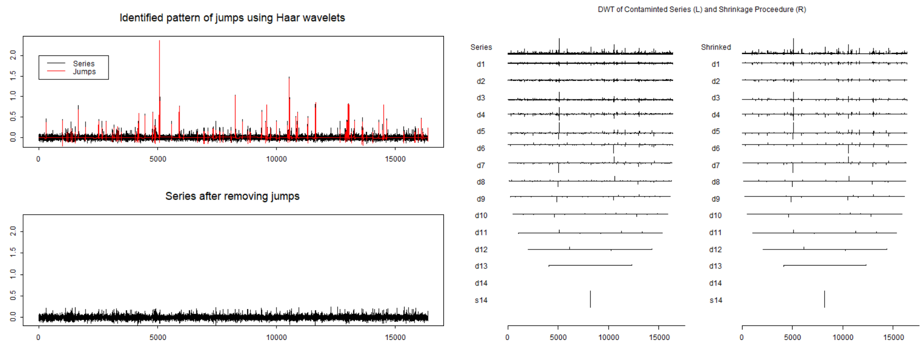
Fig. 2. FIGARCH simulations. (a) Simulated FIGARCH's sample ACF and (b) simulated FIGARCH's sample ACF with jumps.

data, in Fig. 2(b), the long-memory effect on the variance is not as strong because the behavior of the jumps dominates the squared series. Notably, there are just a few statistically non-null lags at a 5% significance level in the sample PACF of series dominated by the jumps.

The result of the GPH Test[20] performed on $Y_t^2$ ($\hat{d} = 0.1271$ with test statistics 2.0966) suggests the presence of long-range dependence at a 5% significance level, whereas the GPH Test performed in $Z_t^2$ ($\hat{d} = -0.0142$ with test statistics $-0.2348$) suggests no evidence of long-range dependence at a 5% significance level.

Therefore, the model could be misidentified if the jumps were not separated out before fitting the model. Therefore, following the heuristics indicated in the proposed methodology, we chose an ARFIMA(1, 1) - GARCH(1, 2) model for the $Z_t$ series, because it was the best in terms of the BIC criterion. In addition, we also fit an ARFIMA(1, 1)-FIGARCH(1, 1) to allow a comparative analysis.

Figure 3(a) shows the identified pattern of jumps using the DTWT and the series after separating out the jumps, whereas Fig. 3(b) shows the shrinkage procedure in detail. After that procedure, we then fit a FIGARCH(1, 1) model to $\hat{Y}_t$, because of the dependence structure seen in the ACF and PACF of the $\hat{Y}_t$ series in Fig. 4.




(a) (b)

Fig. 3. Identification and remotion of jumps with wavelets. (a) Identified pattern of jumps using wavelets and (b) shrinkage procedure.

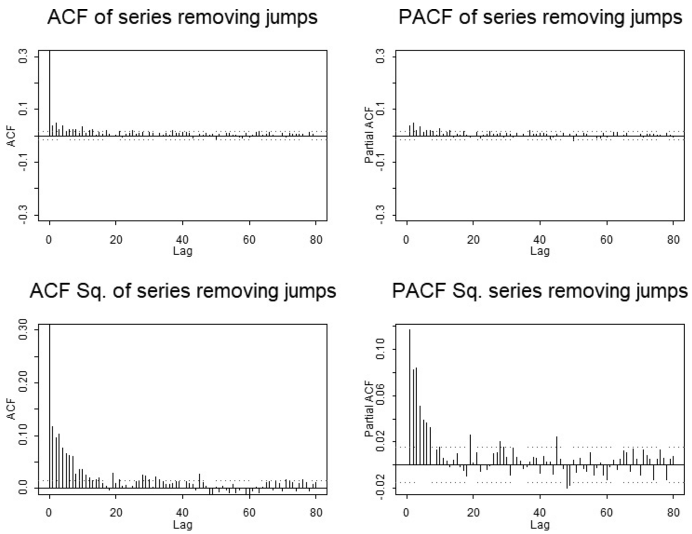



Fig. 4. ACF and PACF of the estimated series after separating out the jumps.

Moreover, the Ljung-Box tests performed on the standardized residuals and squared standardized residuals showed no statistical significant dependence at the 5% level. We also fit an INGARCH(1, 1) model for the estimated jump series obtained from the wavelet shrinkage procedure

When comparing the RMSE of the models, as in Eqs. (22) and (23), we have 389.44 for the ARFIMA$(1, 1)$-GARCH$(1, 1)$ model without the usage of wavelets, and 625.87 for the ARFIMA$(1, 1)$-FIGARCH$(1, 1)$ model without separating out the jumps. However, when we use the wavelet procedure and fit an ACP$(1, 1)$-FIGARCH$(1, 1)$ to the series, we have an RMSE of 232.06, showing that not separating and modeling the jumps lead to a poorer fit in terms of RMSE.

Figure 5 summarizes the comparison between the RMSE of the two models before separating out the jumps and after separating out the jumps. The boxplot clearly shows that the proposed method can generate a much more accurate model in terms of RMSE than an approach not separating out the jumps and modeling them. Following the methodology, Table 1 shows summary statistics for the RMSE of the three fitted models. Note that the proposed method performs, on average, around 40% better than the approach using an ARFIMA-GARCH model. It also

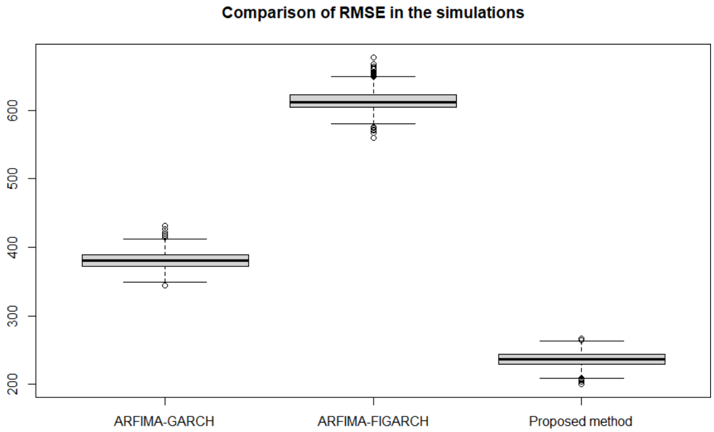



Fig. 5. Boxplot of RMSE of the fitted models.

Table 1. Summary statistics of the RMSE of the models.

| Metrics | ARFIMA-GARCH | ARFIMA-FIGARCH | Proposed method |
|---|---|---|---|
| Mean | 380.720 | 614.452 | 236.036 |
| SD | 12.397 | 14.833 | 10.459 |
| 10% | 365.239 | 598.342 | 221.812 |
| 25% | 372.143 | 604.805 | 229.543 |
| 50% | 380.563 | 612.197 | 236.835 |
| 75% | 388.645 | 622.408 | 243.145 |
| 90% | 396.753 | 634.342 | 248.593 |

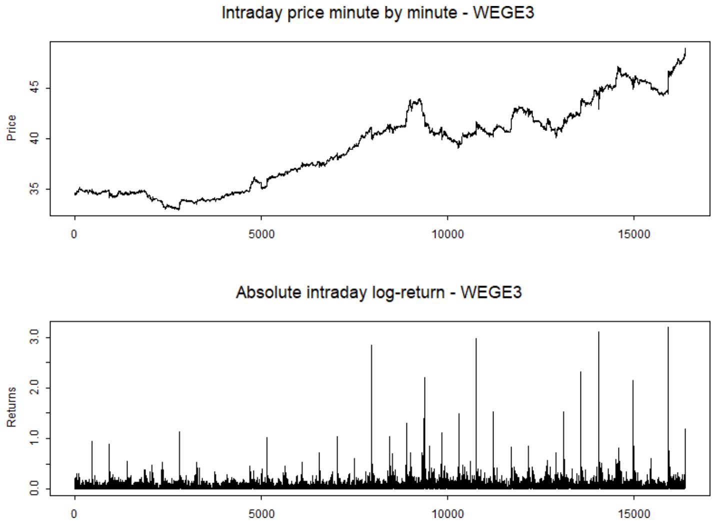



Fig. 6. WEGE3 Intraday price series and absolute log-returns.

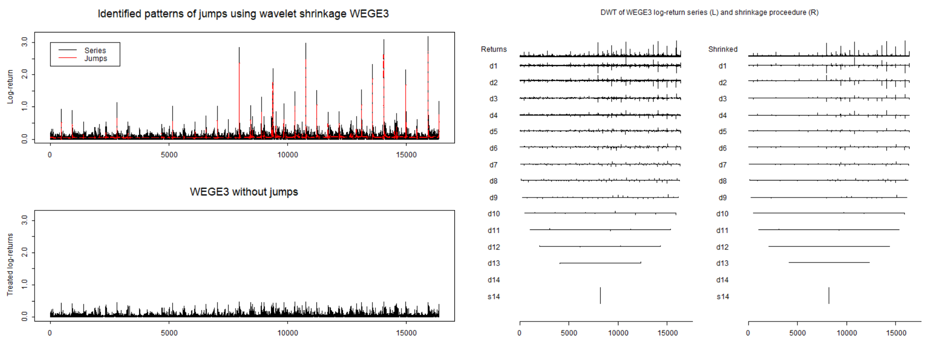
outperforms the ARFIMA-FIGARCH approach with an RMSE that is, on average, three times smaller.

## 5. Empirical Analyses

The data used consists of the minute-by-minute homogeneously sampled time series of absolute log-returns of stocks traded in the Brazilian Stock Exchange and OTC. These were stocks traded on the spot market from January 2, 2020 to February 20, 2020. We also used the ETH/USD cryptocurrency exchange rate traded on the Gemini exchange.

The assets chosen were: FLRY3, WEGE3, and BOVA11. The first three are ordinary shares, and the last one is an exchange-traded fund (ETF). A model was also fitted to the SMLL Index, which acts as a proxy for the behavior of small caps on the Brazilian Stock Exchange. These series were chosen because the stocks are liquid and are from companies in different sectors. Moreover, these series are from different sectors and industries, which means that they might reflect some of their sector's unique characteristics.

The time series chosen to illustrate the procedure is that of the intraday log-returns of WEGE3, homogeneously sampled minute-by-minute. The series starts at 10:03:00 on January 2, 2020 and ends at 10:08:00 on February 20, 2020, which corresponds to approximately one and a half months of trade days (36 days), with $T = 16384$ points. The symbol WEGE3 corresponds to the ordinary shares of WEG S.A., a corporation traded in the Brazilian Exchange and OTC and engaged principally in the capital goods and electrical machines sectors.




(a) (b)

Fig. 7. Jump identification procedure using wavelet shrinkage with MAD function. (a) Identified jump series and (b) DTWT and Shrinakge procedure.

In Fig. 6, note the presence of volatility clusters and jumps in the absolute log-return series, in the sense of anomalous returns from one timestamp to the next.

We calculate the series of log-returns in Fig. 6 using

$$r_t = 100 \times (\ln(P_t) - \ln(P_{t-1})),$$

which means that the largest returns are around 3% from one timestamp to the next. They usually correspond to overnight price changes (19 out of 20).

Figure 7(a) displays the absolute log-returns with the identified pattern of jumps and the series after having the jumps separated out from the original series. The shrinkage function used was the median of the absolute deviation from the median (MAD), in the parameterization proposed in Ref. 6.

Figure 7(b) displays, on the left, the discrete-time wavelet transform of the WEGE3 log-returns using the Haar wavelet and, on the right, the wavelet coefficients after the threshold procedure is applied, using the MAD function.

Next, to obtain the jump series, the inverse discrete-time wavelet transform is applied to the transformed data. In Fig. 7(a), the identified pattern of jumps corresponds to the wavelet reconstruction of the shrunken DTWT procedure of the log-return series as in Fig. 7(b), as a non-parametric estimate.

Figures 8(a) and 8(b) show the ACF and PACF of the log-return series before and after separating out the jumps (anomalous price changes) using the wavelet procedure. Note that, as Ref. 33 describes, such jumps overshadow the long-memory behavior in the series because most of the serial correlation structure lies within the 95% confidence interval, making it statistically insignificant at the 5% significance level. However, after separating out the price jumps, the hidden correlation structure reappears, as shown in Fig. 8(b). The results of the GPH Test on the squared series ($\hat{d} = 0.0264$ with test statistics 0.4354) for long-memory indicate that, at the

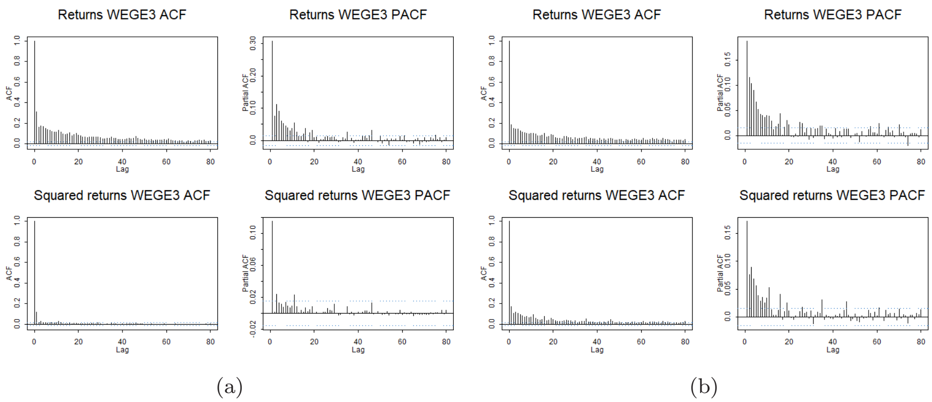



Fig. 8. WEGE3 Sample ACF before and after separating the jumps. (a) Before separating out jumps and (b) after separating out jumps.

1% significance level, there is no statistical evidence of long-memory, contrary to when the jumps are separated out.

An analysis of Fig. 8(b) identifies that the ACF and PACF present the traits of long-range dependence both in the return series and squared return series after separating out the jumps: the ACF has a hyperbolic decay, and there are non-null partial autocorrelations at the 5% significance level in lags close to 30. This means that long-memory models for the conditional mean and conditional variance should be considered.

The ARFIMA$(1, d, 1)$ model was chosen for the conditional mean because it is the most parsimonious while still ensuring that no significant autocorrelations remain at the 5% significance level. The parameters for the ARFIMA model and other helpful information are provided in Table 2.

The model was estimated using the maximum likelihood procedure, using the method of moments estimator for the AR and MA parts as starting values. All the parameters are statistically significant at the 5% level. The estimated $d = 0.3039$ indicates significant long memory in the series. There is also evidence of strong dependence in the short memory part of the model, particularly in the first lag for both the AR and MA parts. The estimated model is stationary since the roots of the AR, and MA polynomials lie outside the unit circle and $|d| < 0.5$. The fact that $d > 0$ means that the model is persistent, which is consistent with the stylized facts of financial time series.

Table 2. ARFIMA coefficients — WEGE3.

| Coef. | Value | Std. Error | $T$ stat. | $p$-value |
|---|---|---|---|---|
| d | 0.3039 | 0.0141 | 21.5817 | 0.0000 |
| AR(1) | 0.3200 | 0.0201 | 15.9326 | 0.0000 |
| MA(1) | 0.5042 | 0.0098 | 51.2308 | 0.0000 |

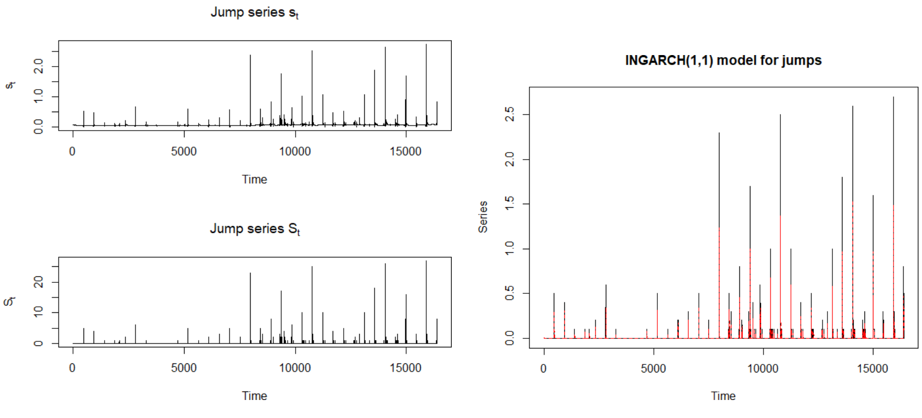
Table 3. FIGARCH coefficients — WEGE3.

| Coef. | Value | Std. Error | $T$ stat. | $p$-value |
|---|---|---|---|---|
| Cst x $10^4$ | 0.4202 | 0.1573 | 2.6710 | 0.0076 |
| d-Figarch | 0.2440 | 0.0421 | 5.7950 | 0.0000 |
| ARCH(1) | 0.3786 | 0.0825 | 4.5900 | 0.0000 |
| GARCH(1) | 0.4440 | 0.0883 | 5.0290 | 0.0000 |
| G.E.D.(DF) | 1.0817 | 0.0194 | 55.8700 | 0.0000 |

We consider an FIGARCH$(1, d, 1)$ model with innovations generated by a generalized error distribution to model the conditional variance model. This model is chosen because it is the most parsimonious while ensuring that there are no significant autocorrelations in the squared residuals of the ARFIMA model at the 5% significance level. Table 3 shows the coefficients for the FIGARCH$(1, d, 1)$ model and the estimated parameter for the GED distribution.

The results summarized in Table 3 lead us to several points. First, the long-memory parameter of the conditional variance ($d = 0.2440$) is smaller than the long-memory parameter of the conditional mean. However, this does not indicate that there is a weaker dependence. As Ref. 7 stressed, the $d$ parameter operates differently for the FIGARCH process: as $d \to 0$, the process's memory increases due to the representation of the effect of a random shock. Given the estimated values in Table 3, the coefficients meet the positivity constraint for the FIGARCH$(1, d, 1)$.[4]

In Fig. 9(a), we can see that the jump series displays some of the traits mentioned before: after the transformation $S_t = \lceil 10 \times s_t \rceil$, there is a significant probability mass at zero; the jumps occur in irregularly spaced times with different magnitudes for




(a) (b)

Fig. 9. WEGE3 Sample ACF before and after separating out the jumps.(a) Identified jumps and (b) fitted INGARCH.

Table 4. INGARCH coefficients — WEGE3.

| Coef. | Value | Std. Error | $T$ stat. | $p$-value |
|---|---|---|---|---|
| Beta(0) | 0.0063 | 0.0006 | 10.9056 | 0.000 |
| Beta(1) | 0.5376 | 0.0272 | 19.7600 | 0.000 |
| Alpha(1) | 0.3729 | 0.0249 | 14.9455 | 0.000 |



every jump. Therefore, it is coherent to fit an Autoregressive Conditional Poisson model for this time series.

In order to model the identified price jumps, an INGARCH$(1, 1)$ model is used with a Poisson distribution. The order is chosen so that there are no remaining statistically significant autocorrelations in the Pearson standardized residuals as a standard procedure.[30] As proposed and implemented in Ref. 30, the coefficients are estimated by quasi-maximum likelihood.

Table 4 shows that all coefficients are statistically significant at the 5% level. From coefficients $\alpha_1$ and $\beta_1$, it is seen that serial dependence in the jump series could be modeled. Moreover, $\alpha_1$ and $\beta_1$ meet the stationarity conditions in Ref. 30 for the INGARCH$(1, 1)$. Figure 9 shows the fitted values for the price jumps versus the observed series. The model successfully describes the pattern of occurrence of the jumps, and the standardized Pearson residuals behave statistically like white noise at the 5% level.

The RMSE of the model after separating the price jumps with wavelets calculated by equation (23) is 0.03461. Given the serial correlation structure exhibited in Fig. 8(a), one can fit an ARFIMA$(3, d, 0)$-GARCH$(1, 1)$ model using the GED distribution, which produces an RMSE of 0.15105. Therefore, the model separating out jumps using wavelets is significantly more accurate in describing the volatility of the intraday series in terms of RMSE.

Table 5 summarizes the RMSE (1) for the models following the proposed methodology, and (2) for the models not separating out the jumps, for all series. It is very clear from these results that the wavelet procedure combined with the long-memory models is much more accurate than the alternative of not removing these jumps in terms of RMSE. Furthermore, it is worth mentioning that in all series, there was no statistical evidence of long memory in the conditional variance before removing jumps, whereas the long-range dependence reappeared in the conditional variance after removing them.

Table 5. Comparison of RMSE of the fitted models.

| Symbol | Proposed method | Not removing jumps | Dif. (%) |
|---|---|---|---|
| WEGE3 | 0.03461 | 0.15105 | −77.09% |
| BOVA11 | 0.06108 | 0.09889 | −38.23% |
| SMLL | 0.0053 | 0.00794 | −33.25% |
| ETH/USD | 0.05514 | 0.07810 | −29.40% |

## 6. Conclusions and Further Remarks

The results indicate that the proposed methodology generates a more accurate estimate of the volatility in terms of the RMSE, which is consistent with the literature. However, a consequence of the bias in $\hat{\rho}(j)$ resulting from not removing the jumps is that the models identified and estimated according to this procedure will result in a poorer fit to the data because they will not capture the long-range dependence of the volatility. This is clear when comparing the RMSE values of the models removing the jumps with wavelets or not removing them, as shown in Table 5. Moreover, using the INGARCH model of the jump series is helpful in describing the process and understanding the dynamics of the jump series.

Volatility is a critical variable in risk management and theoretical economics studies, and accuracy in its estimation is fundamental. The approach using wavelets is interesting because, while it leads to a much more accurate estimation, it does not rely on any subjectiveness or arbitrariness in identifying the jumps, as identification is based on a mathematically objective shrinkage procedure.

Further studies are needed to analyze the behavior of this identified pattern of jumps. Since these price jumps seem not to happen at regular time intervals, the magnitude of the jumps needs to be understood, but also the waiting time between two jumps can be understood as a random variable and thus could also be modeled. Other techniques could be used to study the said jump series to compare how accurately different approaches model them. This is a topic for further research.

## Acknowledgments

In this paper, we used the R programming language,[36] the rugarch package,[21] the tscount package,[30] OxMetrics-G@RCH™ distributed by Timberlake Consultants and described in details with software implementation in Ref. 29 and the S modules S+FinMetrics and S+Wavelets, which were distributed by Insightful Co. and currently by Tibco Spotfire S-PLUS®.

The datasets of minute-by-minute high-frequency time series of the stocks and ETF traded in Brazilian Exchange and OTC, as well as the SMLL Index, were provided by Nelogica,[a] an authorized vendor of the Brazilian Exchange and OTC. Gemini Exchange provides minute-by-minute high-frequency time series for the year 2020, which is then processed and made freely available by CryptoDataDownload.[b] Chiann would like to acknowledge the partial support of FAPESP through the grant 2018/04654-9.

[a]For more information, see Nelogica's web-page: https://www.nelogica.com.br/.
[b]For more information, see CryptoDataDownload's web-page https://www.cryptodatadownload.com/.

## References

1. R. Baille, T. Bollerslev and H. Mikkelsen, Fractionally integrated generalised autoregressive conditional heteroskedasticity, *J. Econom.* **74** (1996) 3–30.
2. J. Beran, Y. Feng, S. Ghosh and R. Kulik, *Long Memory Process*: *Probabilistic Properties and Statistical Methods* (Springer-Verlag, Germany, 2013).
3. M. V. Berry, Z. V. Lewis and J. F. Nye, On the Weierstrass–Mandelbrot fractal function, *Proc. Roy. Soc. Lond. A. Math. Phys. Sci.* **370**(1743) (1980) 459–484.
4. T. Bollerslev and H. Mikkelsen, Modelling and pricing long memory in stock market volatility, *J. Econom.* **73** (1996) 151–184.
5. P. J. Brockwell and R. A. Davis, *Time Series*: *Theory and Methods,* 2nd edn. (Springer, USA, 2009).
6. A. Bruce and H. Gao, *Applied Wavelet Analysis with S-PLUS* (Springer, USA, 1996).
7. M. Caporin, FIGARCH models: Stationarity, estimation methods and the identification problem, Università Ca'Foscari di Venezia, Italy (2002).
8. B. Catalan and F. J. Trivez, Forecasting volatility in GARCH models with additive outliers, *Quant. Finance* **7**(6) (2007) 591–596.
9. W.-S. Chan, Understanding the effect of time series outliers on sample autocorrelations, *Test* **14** (1995) 179–186.
10. I. Chang, Outliers in time series, PhD thesis, University of Wisconsin (1982).
11. I. Chang, G. Tiao and C. Chen, Estimation of time series parameters in the presence of outliers, *Technometrics* **30**(05) (1988) 193–204.
12. A. Charles, Forecasting volatility with outliers in GARCH models, *J. Forecast.* **27**(7) (2008) 551–565.
13. R. Cont, Empirical properties of asset returns: Stylized facts and statistical issues, *Quant. Finance* **1**(03) (2002) 223–236.
14. M. M. Dacorogna, R. Gencay, U. Müller and R. B. Olson, *An Introduction to High-Frequency Finance* (Academic Press, USA, 2001).
15. I. Daubechies, *Ten Lectures of Wavelets* (Springer-Verlag, USA, 1992).
16. T. Downie, Accurate signal estimation near discontinuities, *Int. J. Wavelets Multiresolut. Inform. Process.* **02**(11) (2011) 433–453.
17. R. Ferland, A. Latour and D. Oraichi, Integer-valued GARCH process, *J. Time Ser. Anal.* **27**(6) (2006) 923–942.
18. K. Fokianos, A. Rahbek and D. Tjøstheim, Poisson autoregression, *J. Amer. Statist. Assoc.* **104**(488) (2009) 1430–1439.
19. C. Francq and J. M. Zakoian, *GARCH Models*: *Structure*, *Statistical Inference and Financial Applications* (Wiley and Sons, USA, 2019).
20. J. Geweke and S. Porter-Houdak, The estimation and application of long memory time series models, *J. Time Ser. Anal.* **4** (1983) 221–238.
21. A. Ghalanos, rugarch: Univariate GARCH models. USA, R package version 1.4-2 (2020).
22. E. Guariglia, Entropy and fractal antennas, *Entropy* **18**(3) (2016).
23. E. Guariglia, Harmonic Sierpinski gasket and applications, *Entropy* **20**(9) (2018).
24. E. Guariglia, Primality, fractality, and image analysis, *Entropy* **21**(3) (2019).
25. E. Guariglia and S. Silvestrov, Fractional-wavelet analysis of positive definite distributions and wavelets on $'(\mathbb{C})$, in *Engineering Mathematics II*, eds. S. Silvestrov and M. Rančić (Springer International Publishing, Cham, 2016), pp. 337–353.
26. I. Guttman and G. Tiao, Effect of correlation on the estimation of a mean in the presence of spurious observations, *La Revye Canadienne de Statistique* **6**(12) (1978) 229–247.

27. L. Hotta, The effect of additive outliers on the estimates from aggregated and disaggregated arima models, *Int. J. Forecast.* **9**(1) (1993) 85–93.
28. L. K. Hotta and M. Zevallos, Test of outliers and influential observations in GARCH models: A review, *Estad istica* **64**(10) (2013) 99–119.
29. S. Laurent, *Estimating and Forecasting ARCH Models Using G@RCH$^{TM}$8*: *Ox Metrics* (Timberlake Consultants, United Kingdom, 2018).
30. T. Liboschik, K. Fokianos and R. Fried, tscount: An R package for analysis of count time series following generalized linear models, *J. Statist. Softw.* **82**(5) (2017) 1–51.
31. S. R. C. Lopes and T. S. Prass, Theoretical results on fractionally integrated exponential generalized autoregressive conditional heteroskediastic process, *Physica A: Statist. Mech. Appl.* **401** (2013) 278–307.
32. S. Mallat, *A Wavelet Tour of Signal Processing*: *The Sparse Way* (Elsevier, USA, 2009).
33. B. V. d. M. Mendes, Assessing the bias of maximum likelihood estimates of contaminated garch models, *J. Statist. Comput. Simul.* **67**(11) (2000) 359–376.
34. Y. Meyer, *Wavelets*: *Algorithms and Applications* (SIAM, USA, 1993).
35. G. P. Nason, *Wavelet Methods in Statistics with R* (*Use R!*) (Springer, United Kingdom, 2008).
36. R Core Team, *R*: *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).
37. J. Russell and R. Engle, Analysis of high-frequency data, in *The Handbook of Financial Econometrics*: *Tools and Techniques*, Vol. 1 (Elsevier Science, USA, 2010), pp. 383–426.
38. F. Sowell, Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *J. Econom.* **53**(1) (1992) 165–188.
39. R. S. Tsay, *Analysis of Financial Time Series*, 2nd edn. (Wiley and Sons, USA, 2005).
40. R. S. Tsay, Outliers, level shifts, and variance changes in time series, *J. Forecast.* **7** (1988) 1–20.
41. B. Vidakovic, *Statistical Modeling by Wavelets* (Wiley and Sons, USA, 1999).
42. L. Yang, H. Su, C. Zhong, Z. Meng, H. Luo, X. Li, Y. Y. Tang and Y. Lu, Hyperspectral image classification using wavelet transform-based smooth ordering, *Int. J. Wavelets Multiresolut. Inform. Process.* **17**(06) (2019) 1950050.
43. X. Zheng, Y. Y. Tang and J. Zhou, A framework of adaptive multiscale wavelet decomposition for signals on undirected graphs, *IEEE Trans. Signal Process.* **67**(7) (2019) 1696–1711.
44. E. Zivot and J. Wang, *Modelling Financial Time Series with S-PLUS*, 2nd edn. (Springer, USA, 2005).